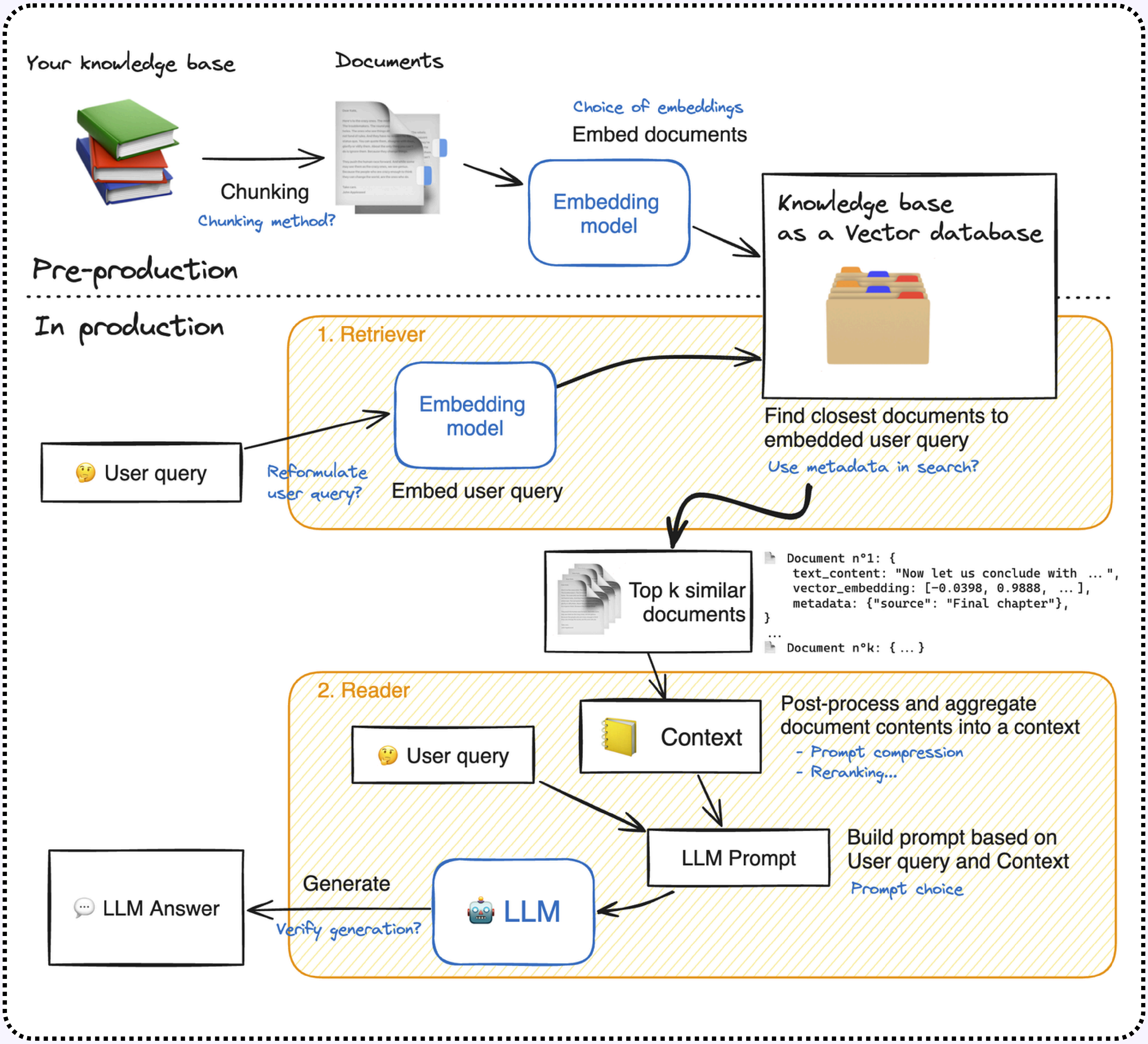
©Harshit Ahluwalia

A Perfect Guide to RAG Evaluation



```
Document n°1: {
  text_content: "Now let us conclude with …",
  vector_embedding: [-0.0398, 0.9888, …],
  metadata: {"source": "Final chapter"},
}
 ...
Document n°k: { … }
```

Types of RAG Evaluation Methods

Retrieval Evaluation Methods

Relevance-Based Evaluation

- Precision@K
- Recall@K
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (NDCG)

Similarity-Based Evaluation

- BM25
- Dense Passage Retrieval (DPR) Similarity Score

Generation Evaluation Methods

Automatic Text Evaluation Metrics

- BLEU (Bilingual Evaluation Understudy)
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
- METEOR (Metric for Evaluation of Translation with Explicit ORdering)
- BERTScore

Hallucination and Factual Consistency Checks

- Faithfulness Evaluation
- Knowledge Grounding Score

Human Evaluation Methods

Relevance and Coherence Assessment

- Likert Scale Ratings
- Pairwise Comparison
- Error Categorization

Domain-Specific Human Evaluation

- Expert Review
- Real-World Application Testing

End-to-End Performance Evaluation

Task-Specific Metrics

- Question Answering Accuracy
- Summarization Quality

User-Centric Metrics

- User Engagement Rate
- A/B Testing

Latency and Efficiency Metrics

- Response Time
- Computational Cost Analysis

©Harshit Ahluwalia

Retrieval Evaluation Methods

Since RAG relies on retrieving relevant documents before generating responses, the retrieval component's effectiveness is crucial.

Relevance-Based Evaluation

- **Precision@K** – Measures the proportion of relevant documents among the top K retrieved.
- **Recall@K** – Evaluates how many relevant documents were retrieved out of all available relevant documents.
- **Mean Reciprocal Rank (MRR)** – Assesses how early the first relevant document appears in the ranked list.
- **Normalized Discounted Cumulative Gain (NDCG)** – Prioritizes highly relevant documents appearing earlier in the ranking.

Similarity-Based Evaluation

- **BM25** – A ranking function to evaluate term-based relevance.
- **Dense Passage Retrieval (DPR) Similarity Score** – Measures how semantically close retrieved passages are to the query using dense embeddings.

©Harshit Ahluwalia

Generation Evaluation Methods

The quality of the generated response is just as critical as retrieval. Various automatic and human evaluations are employed.

Automatic Text Evaluation Metrics

- **BLEU (Bilingual Evaluation Understudy)** – Measures n-gram overlap between generated text and reference text.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** – Focuses on recall, making it useful for summarization tasks.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** – Considers synonyms, stemming, and word order.
- **BERTScore** – Uses contextual embeddings from BERT to compare the similarity between generated and reference texts.

©Harshit Ahluwalia

Hallucination and Factual Consistency Checks

- **Fact-Checking Models (e.g., FEVER dataset)** – Validate the response against known facts.
- **Faithfulness Evaluation** – Measures how well the generated text aligns with the retrieved sources.
- **Knowledge Grounding Score** – Ensures that generated text is derived directly from the retrieved documents.

Human Evaluation Methods

Although automated metrics are useful, human assessment is necessary for nuanced understanding.

Relevance and Coherence Assessment

- **Likert Scale Ratings** – Human judges rate the output on a scale (e.g., 1-5) for relevance, coherence, and fluency.
- **Pairwise Comparison** – Humans compare different RAG outputs to choose the most appropriate one.
- **Error Categorization** – Annotators identify specific errors (e.g., factual inaccuracies, redundancy, or irrelevance).

©Harshit Ahluwalia

Domain-Specific Human Evaluation

- **Expert Review** – Specialists in law, healthcare, or finance evaluate responses based on domain knowledge.

- **Real-World Application Testing** – Evaluates RAG models in production by measuring user engagement, correctness, and usability.

End-to-End Performance Evaluation

A holistic evaluation approach considering real-world applications.

Task-Specific Metrics

- **Question Answering Accuracy** – Measures how well the model answers user queries.

- **Summarization Quality** – Evaluates generated summaries based on brevity, coherence, and informativeness.

©Harshit Ahluwalia

User-Centric Metrics

- **User Engagement Rate** – Measures how often users interact with or accept the model’s responses.

- **A/B Testing** – Compares different RAG configurations to determine the most effective setup.

Latency and Efficiency Metrics

- **Response Time** – Measures how quickly the model generates a response.

- **Computational Cost Analysis** – Evaluates GPU/CPU resource usage for scalability considerations.

Moreover,
we are offering a

Free Certification

on RAG, check the link in the description

@Harshit Ahluwalia